Appendix C. Source and Accuracy Statement

SOURCE OF DATA

The SIPP universe is the noninstitutionalized resident population living in the United States. This population includes persons living in group quarters, such as dormitories, rooming houses, and religious group dwellings. Crew members of merchant vessels, Armed Forces personnel living in military barracks, and institutionalized persons, such as correctional facility inmates and nursing home residents, were not eligible to be in the survey. Also, United States citizens residing abroad were not eligible to be in the survey. Foreign visitors who work or attend school in this country and their families were eligible; all others were not eligible. With the exceptions noted above, persons who were at least 15 years of age at the time of the interview were eligible to be interviewed in the survey.

The 1987 data was collected during the seventh wave of the 1985 panel and the fourth wave of the 1986 panel. Comparisons are made with 1984 data collected during the fourth wave of the 1984 panel.

The 1985 and 1986 panel SIPP samples are located in 230 Primary Sampling Units (PSUs) each consisting of a county or a group of contiguous counties. Within these PSUs, expected clusters of two to four living quarters (LQs) were systematically selected from lists of addresses prepared for the 1980 decennial census to form the bulk of the sample. To account for LQs built within each of the sample areas after the 1980 census, a sample was drawn of permits issued for construction of residential LQs up until shortly before the beginning of the panels.

The 1984 panel SIPP sample was located in 174 PSUs comprising 450 counties and independent cities. Within those PSUs, the bulk of the sample consisted of clusters of two to four LQs, systematically selected from lists of addresses prepared for the 1970 decennial census. The sample was updated to reflect new construction.

In jurisdictions that do not issue building permits, small land areas were sampled and the LQs within were listed by field personnel and then subsampled. In addition, sample LQs were selected from a supplemental frame that included LQs identified as missed in the 1980 census.

The first interviews were conducted during February, March, April, and May of the panel year. Approximately one-fourth of the sample was interviewed in each of these months. Each sample person was visited every 4 months thereafter. At each interview the reference period was the 4 months preceding the interview month.

For subsequent interviews, only original sample persons, those in Wave 1 sample households and interviewed in Wave 1 (and/or 2 for 1985 panel), and persons living with them were eligible to be interviewed. Original sample persons were followed if they moved to a new address, unless the new address was more than 100 miles from a SIPP sample area. Then, telephone interviews were attempted. All first-interview noninterviewed households were automatically designated as noninterviews for all subsequent interviews. When original sample persons moved to remote parts of the country, moved without leaving a forwarding address or refused to be interviewed, additional noninterviews resulted.

As a part of most waves, subjects are covered that do not require repeated measurement during the panel and are of particular interest cross-sectionally for research purposes. A specific set of topical questions are referred to as a topical module. For this report the topical modules analyzed include questions on Retirement and Pension Coverage. They were implemented in Wave 7 of the 1985 panel and Wave 4 of the 1986 panel.

Since Wave 7 of the 1985 panel and Wave 4 of the 1986 panel are concurrent and contain the same relevant topical modules on Retirement and Pension Coverage, the data were combined and analyzed as a single data set. The primary motivation for combining this data is to obtain an increase in sample size in conjunction with a possible reduction in time in sample bias, if any, due to nonresponse over the life of the panel.

Noninterviews. The 1987 tabulations in this report were drawn from interviews conducted from January through April of 1987. Table C-1 summarizes information on nonresponse for the interview months in which the data used to produce this report were collected.

The 1984 estimates are drawn from interviews conducted from September through December 1984. Table C-2 summarizes information on nonresponse for those interview months.

Some respondents do not respond to some of the questions. Therefore, the overall nonresponse rate for some items such as income and money related items is higher than the nonresponse rates in table C-1 and C-2.

Table C-1. **Household Sample Size by Month and Interview Status for 1987 Estimates (1985 and 1986 panels combined)**

Month	Eligible	Inter-viewed	Non inter-viewed	Nonre-sponse rate (percent)[1]
Jan. 1987	6,500	5,300	1,100	17
Feb. 1987	6,700	5,400	1,300	19
Mar. 1987	6,700	5,400	1,300	19
Apr. 1987	6,600	5,300	1,300	19

[1]Due to rounding of all numbers at 100, there are some inconsistencies. The percentage was calculated using unrounded numbers.

Table C-2. **Household Sample Size by Month and Interview Status for 1984 Estimates (1984 panel)**

Month	Eligible	Inter-viewed	Non-inter-viewed	Nonre-sponse rate (percent)[1]
Sep. 1984	5,600	4,800	800	14
Oct. 1984	5,600	4,800	800	15
Nov. 1984	5,600	4,700	900	15
Dec. 1984	5,600	4,700	900	17

[1]Due to rounding of all numbers at 100, there are some inconsistencies. The percentage was calculated using unrounded numbers.

ESTIMATION

The estimation procedure used to derive SIPP person weights in each panel involved several stages of weight adjustments. In the first wave, each person received a base weight equal to the inverse of his/her probability of selection. For each subsequent interview, each person received a base weight that accounted for following movers.

A noninterview factor was applied to the weight of every occupant of interviewed households to account for persons in noninterviewed occupied households which were eligible for the sample. (Individual nonresponse within partially interviewed households was treated with imputation. No special adjustment was made for noninterviews in group quarters.) A factor was applied to each interviewed person's weight to account for the SIPP sample areas not having the same population distribution as the strata from which they were selected. The Bureau has used complex techniques to adjust the weights for nonresponse, but the success of these techniques in avoiding bias is unknown.

An additional stage of adjustment to persons' weights was performed to reduce the mean square errors of the survey estimates. This was accomplished by bringing the sample estimates into agreement with monthly Current Population Survey (CPS) type estimates of the civilian (and some military) noninstitutional population of the United States by demographic characteristics including age, sex, race, and Hispanic ethnicity as of the specified date. The CPS estimates by age, race, sex, and Hispanic origin were themselves brought into agreement with estimates from the 1980 decennial census which have been adjusted to reflect births, deaths, immigration, emigration, and changes in the Armed Forces since 1980. Also, an adjustment was made so that husbands and wives within the same household were assigned equal weights. All of the above adjustments are implemented for each reference month and the interview month.

ACCURACY OF ESTIMATES

SIPP estimates are based on a sample; they may differ somewhat from the figures that would have been obtained if a complete census had been taken using the same questionnaire, instructions, and enumerators. There are two types of errors possible in an estimate based on a sample survey: nonsampling and sampling. We are able to provide estimates of the magnitude of SIPP sampling error, but this is not true of nonsampling error. Found in the next sections are descriptions of sources of SIPP nonsampling error, followed by a discussion of sampling error, its estimation, and its use in data analysis.

Nonsampling variability. Nonsampling errors can be attributed to many sources, e.g., inability to obtain information about all cases in the sample, definitional difficulties, differences in the interpretation of questions, inability or unwillingness on the part of the respondents to provide correct information, inability to recall information, errors made in collection such as in recording or coding the data, errors made in processing the data, errors made in estimating values for missing data, biases resulting from the differing recall periods caused by the interviewing pattern used, and failure of all units in the universe to have some probability of being selected for the sample (undercoverage). Quality control and edit procedures were used to reduce errors made by respondents, coders and interviewers. More detailed discussions of the existence and control of nonsampling errors in the SIPP can be found in the *Quality Profile for the Survey of Income and Program Participation,* May 1990, by Jabine assisted by King and Petroni.

Undercoverage in SIPP results from missed living quarters and missed persons within sample households. It is known that undercoverage varies with age, race, and sex. Generally, undercoverage is larger for males than for females and larger for Blacks than for non-Blacks. Ratio estimation to independent age-race-sex population controls partially corrects for the bias due to survey undercoverage. However, biases exist in the estimates to the extent that persons in missed households or missed persons in interviewed households have characteristics different from those of interviewed persons in the same age-race-sex group. Further, the independent population controls used have not been adjusted for undercoverage.

Unique to the 1986 Panel, maximum telephone interviewing was tested in Waves 2, 3, and 4. Specifically, half of the sample in rotations 4 and 1 of Wave 2 and rotations 2 and 3 of Wave 3 (Phase I) and rotations 2, 3, and 4 of Wave 4 (Phase II) were designated for telephone interviews. Analysis (done by designated mode) of household nonresponse, item nonresponse rates for labor force and income core items, and selected cross-sectional estimates of recipiency, income, low income status, and selected topical module items gave no indication of an overall significant mode effect. However, analysis was restricted to a limited number and type of estimates. If differences between two time periods or differences in characteristics for demographic groups result in borderline significant differences, the significance may be due to bias from the use of the telephone mode. Similarly, borderline insignificant differences may also be due to this bias. Thus, although no overall significant mode effect was detected, the user should consider the possibility of mode effects while analyzing exclusively the 1986 Panel data or combined data involving the 1986 Panel after Wave 1, especially results based on Waves 2 through 4 data. Details on analyses are in "Effect of Maximum Telephone Interviewing on SIPP Topical Module and Longitudinal Estimates" (paper by Gbur, Cantwell and Petroni in the forthcoming *1990 Proceedings of the Survey Research Methods Section, American Statistical Association)* and "SIPP 86: User Statement on Preliminary Results of Maximum Telephone Interviewing" (internal Census Bureau memorandum from Waite to Iannelli, June 6, 1990).

Comparability with other estimates. Caution should be exercised when comparing data from this report with data from other SIPP publications or with data from other surveys. The authors of this report compare SIPP and Current Population Survey (CPS) estimates of pension participation in a technical note. The reader should be cautious in using this information due to the comparability problems. The comparability problems are caused by such sources as the seasonal patterns for many characteristics, different nonsampling errors, and different concepts and procedures. Refer to the *SIPP Quality Profile* for known differences with data from other sources and further discussion. Refer to appendix B P-60 series for a description of CPS survey design.

Sampling variability. Standard errors indicate the magnitude of the sampling error. They also partially measure the effect of some nonsampling errors in response and enumeration, but do not measure any systematic biases in the data. The standard errors for the most part measure the variations that occurred by chance because a sample rather than the entire population was surveyed.

USES AND COMPUTATION OF STANDARD ERRORS

Confidence intervals. The sample estimate and its standard error enable one to construct confidence intervals, ranges that would include the average result of all possible samples with a known probability. For example, if all possible samples were selected, each of these being surveyed under essentially the same conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then:

1. Approximately 68 percent of the intervals from one standard error below the estimate to one standard error above the estimate would include the average result of all possible samples.

2. Approximately 90 percent of the intervals from 1.6 standard errors below the estimate to 1.6 standard errors above the estimate would include the average result of all possible samples.

3. Approximately 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the average result of all possible samples.

The average estimate derived from all possible samples is or is not contained in any particular computed interval. However, for a particular sample, one can say with a specified confidence that the average estimate derived from all possible samples is included in the confidence interval.

Hypothesis testing. Standard errors may also be used for hypothesis testing, a procedure for distinguishing between population characteristics using sample estimates. The most common types of hypotheses tested are 1) the population characteristics are identical versus 2) they are different. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical.

All statements of comparison in the report have passed a hypothesis test at the 0.10 level of significance or better. This means that, for differences cited in the report, the estimated absolute difference between parameters is greater than 1.6 times the standard error of the difference.

To perform the most common test, compute the difference $X_A - X_B$, where X_A and X_B are sample estimates of the characteristics of interest. A later section explains how to derive an estimate of the standard error of the difference $X_A - X_B$. Let that standard error be s_{DIFF}. If $X_A - X_B$ is between -1.6 times s_{DIFF} and +1.6 times s_{DIFF}, no conclusion about the characteristics is justified at the 10 percent significance level. If, on the other hand, $X_A - X_B$ is smaller than -1.6 times s_{DIFF} or larger than +1.6 times s_{DIFF}, the observed difference is significant at the 10 percent level. In this event, it is commonly accepted practice to say that the characteristics are different. Of course, sometimes this conclusion will be wrong. When the characteristics are, in fact, the same, there is a 10 percent chance of concluding that they are different.

Note that as more tests are performed, more erroneous significant differences will occur. For example, at the 10 percent significance level, if 100 independent hypothesis tests are performed in which there are no real differences, it is likely that about 10 erroneous differences will occur. Therefore, the significance of any single test should be interpreted cautiously.

Note concerning small estimates and small differences. Summary measures are shown in the report only when the base is 200,000 or greater. Because of the large standard errors involved, there is little chance that estimates will reveal useful information when computed on a base smaller than 200,000. Also, nonsampling error in one or more of the small number of cases providing the estimate can cause large relative error in that particular estimate. Estimated numbers are shown, however, even though the relative standard errors of these numbers are larger than those for the corresponding percentages. These smaller estimates are provided primarily to permit such combinations of the categories as serve each user's needs. Therefore, care must be taken in the interpretation of small differences since even a small amount of nonsampling error can cause a borderline difference to appear significant or not, thus distorting a seemingly valid hypothesis test.

Standard error parameters and tables and their use. Most SIPP estimates have greater standard errors than those obtained through a simple random sample because clusters of living quarters are sampled for the SIPP. To derive standard errors that would be applicable to a wide variety of estimates and could be prepared at a moderate cost, a number of approximations were required. Estimates with similar standard error behavior were grouped together and two parameters (denoted "a" and "b") were developed to approximate the standard error behavior of each group of estimates. Because the actual standard error behavior was not identical for all estimates within a group, the standard errors computed from these parameters provide an indication of the order of magnitude of the standard error for any specific estimate. These "a" and "b" parameters vary by characteristic and by demographic subgroup to which the estimate applies. Tables C-8 and C-9 provide base "a" and "b" parameters to be used for the 1984 estimates and the 1987 estimates, respectively.

For those users who wish further simplification, we have also provided general standard errors in tables C-4 through C-7. Note that these standard errors must be adjusted by a factor from tables C-8 and C-9. The standard errors resulting from this simplified approach are less accurate. Methods for using these parameters and tables for computation of standard errors are given in the following sections.

Standard errors of estimated numbers. There are two ways to compute the approximate standard error, s_x, of an estimated number shown in this report. The first uses the formula

$$s_x = fs \qquad (1)$$

where f is a factor from table C-8 or C-9, and s is the standard error of the estimate obtained by interpolation from table C-4 or C-5. Alternatively, s_x may be approximated by the formula,

$$s_x = \sqrt{ax^2 + bx} \qquad (2)$$

from which the standard errors in tables C-4 or C-5 were calculated. Here x is the size of the estimate and a and b are the parameters in tables C-8 or C-9 associated with the particular type of characteristic. Use of formula 2 will provide more accurate results than the use of formula 1. When calculating standard errors for numbers from cross-tabulations involving different characteristics, use the factor or set of parameters for the characteristic which will give the largest standard error.

Table C-3. **Monthly Earnings of Worker with IRAs, 1987 Estimates**

Level of earnings	Number (in thousands)	Percent with at least as much as lower bound of interval
Total	20,033	-
Under $1,000	3,178	100.00
1,000 to 1,499	3,075	84.1
1,500 to 1,999	3,320	68.8
2,000 to 2,499	2,872	52.2
2,500 to 2,999	2,240	37.9
3,000 to 3,499	1,826	26.7
3,500 to 3,999	933	17.6
4,000 and over	2,590	12.9

-Represents zero.

Table C-4. **Standard Errors of Estimated Numbers of Persons, 1984 Estimates**

(Numbers in thousands)

Size of estimate	Standard error	Size of estimate	Standard error
200	63	23,000	641
300	77	25,000	666
600	109	27,000	689
1,000	141	30,000	721
2,000	199	50,000	883
5,000	312	80,000	1,020
8,000	392	100,000	1,062
11,000	457	130,000	1,062
13,000	494	150,000	1,021
15,000	528	170,000	937
17,000	560	200,000	725
20,000	601		

Table C-5. **Standard Errors of Estimated Numbers of Persons, 1987 Estimates**

(Numbers in thousands)

Size of estimate	Standard error	Size of estimate	Standard error
200	59	23,000	597
300	72	25,000	620
600	102	27,000	641
1,000	131	30,000	670
2,000	185	50,000	821
5,000	290	80,000	949
8,000	365	100,000	988
11,000	425	130,000	988
13,000	460	150,000	950
15,000	491	170,000	877
17,000	521	200,000	676
20,000	561		

Illustration. The number of wage and salary workers covered by pension plans in 1987 is 55,700,000. The appropriate parameters and factor for 1987 come from table C-9 and the appropriate general standard error from table C-5 are

$a = -0.00005085$, $b = 8674$, $f = 0.71$, $s = 845{,}000$

Using formula (1), the approximate standard error is $s_x = 845{,}000 \times 0.71 = 600{,}000$.

Using formula (2), the approximate standard error is

$$s_x = \sqrt{-0.00005085(55{,}700{,}000)^2 + 8{,}674(55{,}700{,}000)} = 570{,}000$$

Using the standard error based on formula (2), the approximate 90-percent confidence interval as shown by the data is from 54,788,000 to 56,612,000.

Standard errors of estimated perçentages. The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends on the size of the percentage and its base. When the numerator and denominator of the percentage have different parameters, use the parameter (or appropriate factor) from tables C-6 through C-9 indicated by the numerator.

The approximate standard error, $s_{(x,p)}$, of an estimated percentage p can be obtained by use of the formula

$$s_{(x,p)} = fs \qquad (3)$$

where p is the percentage of persons/families/households with a particular characteristic such as the percent of workers covered by a pension plan.

In this formula, f is the appropriate "f" factor from tables C-8 or C-9, and s is the standard error of the estimate obtained by interpolation from tables C-6 or C-7. Alternatively, it may be approximated by the formula:

$$s_{(x,p)} = \sqrt{\frac{b}{x}(p)(100-p)} \qquad (4)$$

from which the standard errors in tables C-6 and C-7 were calculated. Here x is the total number of persons, families, households, or unrelated individuals in the base of the percentage, p is the percentage ($0 \leq p \leq 100$), and b is the "b" parameter in tables C-8 and C-9 associated with the characteristic in the numerator of the percentage. Use of this formula will give more accurate results than use of formula (3) above.

Table C-6. **Standard Errors of Estimated Percentages of Persons, 1984 Estimates**

Base of estimated percentage (thousands)	Estimated percentage					
	≤ 1 or ≥ 99	2 or 98	5 or 95	10 or 90	25 or 75	50
200	3.1	4.4	6.9	9.5	13.7	15.8
300	2.6	3.6	5.6	7.7	11.2	12.9
600	1.8	2.6	4.0	5.5	7.9	9.1
1,000	1.4	2.0	3.1	4.2	6.1	7.1
2,000	1.0	1.4	2.2	3.0	4.3	5.0
5,000	0.6	0.9	1.4	1.9	2.7	3.2
8,000	0.5	0.7	1.1	1.5	2.2	2.5
11,000	0.4	0.6	0.9	1.3	1.8	2.1
13,000	0.4	0.5	0.8	1.2	1.7	2.0
17,000	0.34	0.5	0.7	1.0	1.5	1.7
22,000	0.29	0.4	0.7	0.9	1.3	1.5
26,000	0.28	0.4	0.6	0.8	1.2	1.4
30,000	0.26	0.4	0.6	0.8	1.1	1.3
50,000	0.20	0.3	0.4	0.6	0.9	1.0
80,000	0.16	0.2	0.3	0.5	0.7	0.8
100,000	0.14	0.2	0.3	0.4	0.6	0.7
130,000	0.12	0.17	0.3	0.4	0.5	0.6
200,000	0.10	0.13	0.2	0.3	0.4	0.5

Table C-7. **Standard Errors of Estimated Percentages of Persons, 1987 Estimates**

Base of estimated percentage (thousands)	Estimated percentage					
	≤ 1 or ≥ 99	2 or 98	5 or 95	10 or 90	25 or 75	50
200	2.9	4.1	6.4	8.8	12.7	14.7
300	2.4	3.4	5.2	7.2	10.4	11.9
600	1.7	2.4	3.7	5.1	7.3	8.5
1,000	1.3	1.8	2.9	3.9	5.7	6.6
2,000	0.9	1.3	2.0	2.8	4.0	4.6
5,000	0.6	0.8	1.3	1.8	2.5	2.9
8,000	0.5	0.6	1.0	1.4	2.0	2.3
11,000	0.4	0.6	0.9	1.2	1.7	1.9
13,000	0.4	0.5	0.8	1.1	1.6	1.8
15,000	0.3	0.5	0.7	1.0	1.5	1.7
17,000	0.3	0.4	0.7	1.0	1.4	1.6
20,000	0.29	0.4	0.6	0.9	1.3	1.5
23,000	0.27	0.4	0.6	0.8	1.2	1.4
25,000	0.26	0.4	0.6	0.8	1.1	1.3
27,000	0.25	0.3	0.5	0.7	1.1	1.3
30,000	0.23	0.3	0.5	0.7	1.0	1.2
50,000	0.18	0.3	0.4	0.5	0.8	0.9
80,000	0.15	0.2	0.3	0.4	0.6	0.7
100,000	0.13	0.18	0.3	0.4	0.6	0.7
130,000	0.11	0.16	0.2	0.3	0.5	0.6
150,000	0.11	0.15	0.2	0.3	0.5	0.5
170,000	0.10	0.14	0.2	0.3	0.4	0.5
200,000	0.09	0.13	0.2	0.3	0.4	0.5

Illustration. Table A in the report shows that 66.4 percent of workers were covered by a pension plan in 1987. Using formula (3) with the factor from table C-9 and the appropriate standard error from table C-7, the approximate standard error is

$$s_{(x,p)} = .71\text{x}.656\% = 0.47\%$$

Using formula (4) with the "b" parameter from table C-9, the approximate standard error is

$$s_{(x,p)} = \sqrt{\frac{8{,}674}{83{,}962{,}000}(66.4\%)(100\% - 66.4\%)} = 0.48\%$$

Consequently, the approximate 90-percent confidence interval as shown by these data is from 65.6 to 67.2 percent.

Standard error of a median. The median quantity of some item such as income for a given group of persons, families, or households is that quantity such that at least half the group have as much or more and at least half the group have as much or less. The sampling variability of an estimated median depends upon the form of the distribution of the item as well as the size of the group. To calculate standard errors on medians, the procedure described below may be used.

Note that the standard errors for all median values displayed in detailed tables are usually provided immediately next to the medians. However, if the reader desires to calculate standard errors on medians for collapsed groups, the procedure described below may be used. Also note that the medians and their standard errors given in detailed tables will be somewhat different from those calculated using this method since more interval breaks were used than shown.

An approximate method for measuring the reliability of an estimated median is to determine a confidence interval about it. (See the section on sampling variability for a general discussion of confidence intervals.) The following procedure may be used to estimate the 68-percent confidence limits and hence the standard error of a median based on sample data.

1. Determine, using either formula 3 or formula 4, the standard error of an estimate of 50 percent of the group;
2. Add to and subtract from 50 percent the standard error determined in step 1;
3. Using the distribution of the item within the group, calculate the quantity of the item such that the percent of the group owning more is equal to the smaller percentage found in step 2. This quantity will be the upper limit for the 68-percent confidence interval. In a similar fashion, calculate the quantity of the item such that the percent of the group owning more is equal to the larger percentage found in step 2. This quantity will be the lower limit for the 68-percent confidence interval;
4. Divide the difference between the two quantities determined in step 3 by two to obtain the standard error of the median.

To perform step 3, it will be necessary to interpolate. Different methods of interpolation may be used. The most common are simple linear interpolation and Pareto interpolation. The appropriateness of the method depends on the form of the distribution around the median. If density is declining in the area, then we recommend Pareto interpolation. If density is fairly constant in the area, then we recommend linear interpolation. Note, however, that Pareto interpolation can never be used if the interval contains zero or negative measures of the item of interest. Interpolation is used as follows. The quantity of the item such that "p" percent own more is

$$X_{pN} = \exp\left[\left(\mathrm{Ln}\left(\frac{pN}{N_1}\right) / \mathrm{Ln}\left(\frac{N_2}{N_1}\right)\right) \mathrm{Ln}\left(\frac{A_2}{A_1}\right)\right] A_1 \qquad (5)$$

if Pareto interpolation is indicated and

$$X_{pN} = \left[\frac{pN—N_1}{N_2—N_1}(A_2—A_1) + A_1\right] \qquad (6)$$

if linear interpolation is indicated, where N is the size of the group,

A_1 and A_2 are the lower and upper bounds, respectively, of the interval in which X_{pN} falls,
N_1 and N_2 are the estimated number of group members owning more than A_1 and A_2, respectively,
exp refers to the exponential function and
Ln refers to the natural logarithm function.

Illustration. Using text table E, the median monthly earnings amount of workers with an IRA was $2,077. The size of this group of workers was 20,033,000.

1. Using formula (4), the standard error of 50 percent on a base of 20,033,000 is about 0.8 percentage points.
2. Following step 2, two percentages of interest are 49.2 and 50.8.
3. By examining table C-3, we see that the percentage 49.2 falls in the income interval from 2,000 to 2,499. (Since 52.2 percent receive more than 2,000 per month, but only 37.9 percent receive more than 2,499 per month, the quantity that exactly 49.2 percent receive more than must be between 2,000 and 2,499.) Thus $A_1 = 2{,}000$, $A_2 = 2{,}499$, $N_1 = 10{,}457{,}226$ and $N_2 = 7{,}592{,}507$. In this case, we decided to use Pareto interpolation, formula (5).

Therefore, the upper bound of a 68-percent confidence interval for the median is

$$\exp\left[\left(\ln\left(\frac{(0.492)(20{,}033{,}000)}{10{,}457{,}226}\right) / \ln\left(\frac{7{,}592{,}507}{10{,}457{,}226}\right)\right) \cdot \ln\left(\frac{2{,}499}{2{,}000}\right)\right] \cdot 2{,}000 = 2{,}084$$

Also, by examining table C-3, we see that the percentage of 50.8 falls in income interval from 2,000 to 2,499. Thus, $A_1 = 2{,}000$, $A_2 = 2{,}499$, $N_1 = 10{,}457{,}266$, and $N_2 = 7{,}592{,}507$. Again using Pareto interpolation, formula (5), lower bound of a 68-percent confidence interval for the median is

$$\exp\left[\left(\ln\left(\frac{(0.508)(20{,}033{,}000)}{10{,}457{,}226}\right) / \ln\left(\frac{7{,}592{,}507}{10{,}457{,}226}\right)\right) \cdot \ln\left(\frac{2{,}499}{2{,}000}\right)\right] \cdot 2{,}000 = 2{,}038$$

Thus the 68-percent confidence interval on the estimated median is from $2,038 to $2,084. An approximate standard error is

$$\frac{\$2{,}084—\$2{,}038}{2} = \$23$$

Standard error of a difference. The standard error of a difference between two sample estimates, x and y, is approximately equal to

$$s_{(x-y)} = \sqrt{s_x^2 + s_y^2 - 2rs_xs_y} \qquad (7)$$

where s_x and s_y are the standard errors of the estimates x and y and r is the correlation coefficient between the characteristics estimated by x and y. The estimates can be numbers, averages, percents, ratios, etc. Underestimates or overestimates of standard error of differences result if the estimated correlation coefficient is overestimated or underestimated, respectively. In this report, r is assumed to be zero.

Illustration. Again, using text table A, 52,727,000 workers were covered by a pension plan in 1984 and 55,738,000 workers were covered by a pension plan in 1987. The standard error for the numbers are computed using formula (2) to be 604,000 (using 1984 parameter in table C-8) and 570,000 (using 1987 parameter in table C-9), respectively.

Assuming that these two estimates are not correlated, the standard error of the estimated difference of 3,011,000 is

$$s_{x-y} = \sqrt{(604{,}000)^2 + (570{,}000)^2} = 831{,}000$$

The approximate 90-percent confidence interval is from 1,681,000 to 4,340,000. Since this interval does not contain zero, we conclude that the difference is significant at the 10-percent level.

Standard error of a mean. A mean is defined here to be the average quantity of some item (other than persons) per person. For example, the mean could be the average monthly earnings of workers with IRAs. The standard error of such a mean can be approximated by formula (8) below. Because of the approximations used in developing formula (8), an estimate of the standard error of the mean obtained from that formula will generally underestimate the true standard error.

The formula used to estimate the standard error of a mean $\bar{x}$ is

$$s_{\bar{x}} = \sqrt{\left[\frac{b}{y}\right] s^2} \qquad (8)$$

where y is the size of the base, s^2 is the estimated population variance of the item and b is the parameter associated with the particular type of item.

The population variance s^2 may be estimated by the following: We assume x_i is the value of the item for unit i. (Unit could be person, family, or household.) The range of values for the item is divided into c intervals. The upper and lower boundaries of interval j are Z_{j-1}, and Z_j, respectively. Each unit is placed into one of c groups such that $Z_{j-1} < x_i \leq Z_j$.

The estimated population variance, s^2 is given by formula:

$$s^2 = \sum_{j=1}^{c} p_j m_j^2 - \bar{x}^2 \qquad (9)$$

where p_j is the estimated proportion of units in group j, and $m_j = (Z_{j-1} + Z_j) / 2$. The most representative value of the item in group j is assumed to be m_j. If group c is open-ended, i.e., no upper interval boundary exists, then an approximate value for m_c is

$$m_c = \left(\frac{3}{2}\right) Z_{c-1} \qquad (10)$$

The mean, $\bar{x}$, can be obtained using the following formula:

$$\bar{x} = \sum_{j=1}^{c} p_j m_j \qquad (11)$$

Illustration. The distribution of monthly earnings levels of workers with IRAs in 1987 is given in text table E. Using formulas (9), (10), and (11), and the mean monthly earning amount of $2,446, the approximate population variance for all workers with IRAs, s^2 is

$$s^2 = \left(\frac{3{,}178}{20{,}033}\right)(500)^2 + \left(\frac{3{,}075}{20{,}033}\right)(1{,}250)^2 + \ldots + \left(\frac{2{,}590}{20{,}033}\right)(6{,}000)^2 - (2{,}446)^2 = 2{,}669{,}459$$

Using formula (8) and "b" parameter from table C-9, the estimated standard error of a mean $\bar{x}$ is

$$s_{\bar{x}} = \sqrt{\frac{4{,}736}{20{,}033{,}000}(\$2{,}669{,}459)} = \$25$$

Standard error of a ratio. The standard error for the average quantity of persons, families, or households per family or household or for a ratio of means or medians is approximated by formula (12):

$$s_{x/y} = \frac{x}{y}\sqrt{\left(\frac{s_x}{x}\right)^2 + \left(\frac{s_y}{y}\right)^2 - 2r\frac{s_x s_y}{xy}} \qquad (12)$$

Where x and y are the numerator and denominator for the average or the means or medians which form the ratio, r is the correlation coefficient between the characteristics estimated by x and y. Their associated standard errors are s_x and s_y.

Underestimates or overestimates of standard error of ratios result if the estimated correlation coefficient is overestimated or underestimated, respectively. In this report, r is assumed to be zero.

Illustration. We see in detailed table 5, that, in 1987, the mean pension income of retirees with Cost of Living Adjustment (COLA) provisions was $780. This appears to be $321 higher than the mean pension income of retirees with no COLA provision of $459. The ratio of this mean difference to the mean with no COLA results in 0.7 (or 0.7 X 100 = 70%). The standard error for the ratio result is obtained using formula (12) where x = 780 - 459 = 321, y = 459, s_x = 46.8 (using formula (7)) and s_y = 31 (from table 5). The estimated standard error for this ratio is

$$s_{x/y} = \frac{321}{459}\sqrt{\left(\frac{46.8}{321}\right)^2 + \left(\frac{31}{459}\right)^2} = 0.11$$

Consequently, the approximate 90 percent confidence interval is 0.52 to 0.88 (or 52% to 88%).

Table C-8. **SIPP Generalized Variance Parameters for 1984 Panel**

Characteristic[1]	a	b	f
PERSONS			
Total or White			
16+ Income and Labor Force (3)			
Both Sexes	-0.0000321	5,475	0.52
Male	-0.0000677	5,475	0.52
Female	-0.0000612	5,475	0.52
16+ Pénsion Plan[3] (2)			
Both Sexes	-0.0000588	10,027	0.71
Male	-0.0001240	10,027	0.71
Female	-0.0001121	10,027	0.71
All Others[2] (4)			
Both Sexes	-0.0000864	19,911	1.00
Male	-0.0001786	19,911	1.00
Female	-0.0001672	19,911	1.00
Black (1)			
Both Sexes	-0.0002670	7,366	0.61
Male	-0.0005737	7,366	0.61
Female	-0.0004933	7,366	0.61

[1]For cross-tabulations, use the parameters of the characteristic with the smaller number within the parentheses.
[2]Use the "All Others" parameters for retirement tabulations, 0+ program participation, 0+ benefits, 0+ income, and 0+ labor force tabulations, in addition to any other types of tabulations not specifically covered by another characteristic in this table.
[3]Use the "16+ Pension Plan" parameters for pension plan tabulations of persons 16+ in the labor force.

Table C-9. **SIPP Generalized Variance Parameters for 1987 Estimates (Combined 1985 and 1986 panels parameters)**

Characteristic[1]	a	b	f
PERSONS			
Total or White			
16+ Income and Labor Force (3)			
Both Sexes	-0.0000278	4,736	0.52
Male	-0.0000586	4,736	0.52
Female	-0.0000530	4,736	0.52
16+ Pension Plan[3] (2)			
Both Sexes	-0.0000509	8,674	0.71
Male	-0.0001072	8,674	0.71
Female	-0.0000970	8,674	0.71
All Others[2] (4)			
Both Sexes	-0.0000747	17,224	1.00
Male	-0.0001545	17,224	1.00
Female	-0.0001446	17,224	1.00
Black (1)			
Both Sexes	-0.0002310	6,372	0.61
Male	-0.0004963	6,372	0.61
Female	-0.0004319	6,372	0.61

[1]For cross-tabulations, use the parameters of the characteristic with the smaller number within the parentheses.
[2]Use the "All Others" parameters for retirement tabulations, 0+ program participation, 0+ benefits, 0+ income, and 0+ labor force tabulations, in addition to any other types of tabulations not specifically covered by another characteristic in this table.
[3]Use the "16+ Pension Plan" parameters for pension plan tabulations of persons 16+ in the labor force.